Data Clustering and Similarity

Julien Soler[1] [2], **Fabien Tencé**[1], **Laurent Gaubert**[2] and **Cédric Buche**[2]
[1] Virtualys
fabien.tence@virtualys.com

[2] UEB, Lab-STICC, ENIB
{soler, gaubert, buche}@enib.fr

Abstract

In this article, we study the notion of similarity within the context of cluster analysis. We begin by studying different distances commonly used for this task and highlight certain important properties that they might have, such as the use of data distribution or reduced sensitivity to the curse of dimensionality. Then we study inter- and intra-cluster similarities. We identify how the choices made can influence the nature of the clusters.

1 Introduction

Data clustering is an important part of data mining. This technique is used in many fields such as biological data analysis or image segmentation. The aim is to identify groups of data known as clusters, in which the data are similar.

Effective clustering maximizes intra-cluster similarities and minimizes inter-cluster similarities (Chen, Han, and Yu 1996). Before defining inter- and intra-cluster similarities, first we must define the similarity between a data pair. There are many different ways of defining these similarities and depending on the chosen method, the results of the cluster analysis may strongly differ.

So how can we reliably choose one definition over another? The aim of this article is to provide clues in order to answer this question by studying different definitions of similarity. We will study the following elements separately: distance between two pieces of data and inter-and intra-cluster distances.

In section 2, we will discuss the notion of distance between two pieces of data. We shall examine different distances which can be used in the majority of clustering algorithms and their relevance depending on the problem at hand. We shall see in section 3 how the use of different inter/intra-cluster distances can dote the resulting clusters with certain properties. Finally, we will discuss how our study might be of merit for unresolved issues.

2 Data Similarity

Generally speaking, data similarity is evaluated using the notion of distance. In this section, we will define the notion of distance. First, we shall identify the mathematical properties required in clustering. Then we will turn to the way in which the distances can exploit the distribution of the data to be clustered. We will also see that certain metrics are more appropriate in high-dimensional spaces. We shall also see that it is possible to better separate data by expressing the data within another space, thus increasing the contrast between the distances.

2.1 Mathematical Properties

Formally, a distance on a set E (in our case, E will be a vector space) is defined as an application:
$d : \mathrm{E} \times \mathrm{E} \to \mathbb{R}^+$ with the following properties:

- Symmetry: $\forall \mathbf{x}, \mathbf{y} \in \mathrm{E}, d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
- Separation: $\forall \mathbf{x}, \mathbf{y} \in \mathrm{E}, d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$
- Triangular inequality: $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathrm{E},\ d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$

Examples of distances along with their definitions are presented in Table 1.

Distances which correspond to this definition alone are not necessarily the best solution for accurately clustering the data. Indeed, such a distance may not necessarily be stable by translation, for example ($d(\mathbf{x}+\mathbf{a}, \mathbf{y}+\mathbf{a}) \neq d(\mathbf{x}, \mathbf{y})$). For this reason, most of the distances used stem from the notion of the norm (the distance is thus expressed by $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$), with the addition of the property of homogeneity ($\|k\mathbf{u}\| = |k| * \|\mathbf{u}\|$). Moreover, certain norms derive from a scalar product (by $\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$). and therefore have other additional properties. For instance, it is possible to show that the barycenter of N points is the point which minimizes the squares of the distances to these points. This is a very useful property in the field of cluster analysis. The Euclidean distance stems from the L2 norm, itself defined by the usual scalar product. They thus possess all of these properties.

It is nonetheless possible to use functions which do not even meet the criteria of distance definition. This is true for example in the case of Minkowski distances when $p < 1$, which do not satisfy for triangular inequality. It is therefore necessary to check that the chosen algorithm converges without this property.

2.2 Data Distribution

Distances such as Euclidean distance or Minkowski distances are independent of the data they are used to compare.

Manhattan	$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Euclidean	$d(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{n} (x_i - y_i)^2)^{\frac{1}{2}}$
Euclidean, Standardized	$d(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\sigma_i^2})^{\frac{1}{2}}$
Minkowski	$d(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p)^{\frac{1}{p}}, p \geq 1$
Chebyshev	$d(\mathbf{x},\mathbf{y}) = \lim_{p \to \infty} (\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p)^{\frac{1}{p}}$
Mahalanobis	$d(\mathbf{x},\mathbf{y}) = ((\mathbf{x}-\mathbf{y})^T S^{-1} (\mathbf{x}-\mathbf{y}))^{\frac{1}{2}}$ where S denotes the covariance matrix of the dataset.

Table 1: Defining Different Distances

However the scales of the dimensions are not necessarily comparable. Therefore, when dealing with data relating to a person, for example, the units of age and height are not commensurate. By using the Euclidean distance, data clustering does not discriminate between people according to age alone, because an age difference of one year is as big a difference as a height variation of one meter. The standardized Euclidean distance is much more suited, as when considering each dimension it divides its value by its variance. If we look at this in more detail, using the same example, height and age are correlated dimensions. This correlation carries additional information which can be taken into account. The Mahalanobis distance uses the covariance matrix of the data, thus exploiting the correlation and the variance between the data. However, it is possible for one of the dimensions to be proportional to another. This dimension thus becomes redundant. In such cases, the eigenvalues of the covariance matrix are null, and the Mahalanobis distance therefore cannot be calculated. It is possible to conduct a principal component analysis in order to detect these correlated data and to ignore them. It is interesting to note that the standardized Euclidean distance of the data projected in the eigenspace is equal to the Mahalanobis distance.

Figure 1 depicts image clustering conducted with a randomly initialized k-means algorithm with three centroids by using various distances. The input data for each pixel is its coordinates in x and y along with its red, green and blue color (thus a 5-dimensional vector space). For each distance function, the algorithm was run three times and the most convincing result (as chosen by a human) was retained. We can see the effectiveness of Mahalanobis distance and standardized Euclidean distance compared with other distances.

2.3 The Curse of Dimensionality

When the data originate from a high-dimensional space, we face a problem known as the curse of dimensionality. Dimension reduction is possible by conducting a principal component analysis and retaining only the most significant dimensions. However, this method discards some of the information. A behavioral study of the Minkowski distances on high-dimensional spaces (Aggarwal, Hinneburg, and Keim 2001) shows that the p-distances for a high p-value only exacerbates the problem. As we have already explained, the fractional p-distances are not distances in the formal sense; despite this fact, they can be used to accentuate relative contrast between data. Indeed, they tend to group data together and therefore reduce the curse of dimensionality effect. The Manhattan distance has the advantage of both having triangular inequality and offering better data contrast than Euclidean distance. Furthermore, it is interesting to note in Figure 1, that the clustering calculated with the Manhattan distance does not divide the image into three equal parts, as in the cases of the Euclidean and Minkowski distances with $p = 20$. The clustering seems better than any regular p-distance (Figure 1: b., c. and e.). Figure 2 shows the same image clustered using a fractional p-distance (p=0.2).

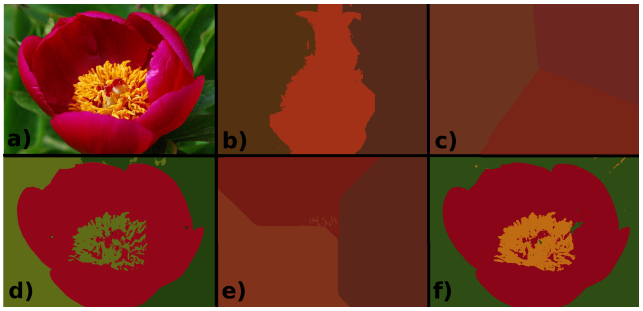

Figure 1: Illustration of the importance of the distance function on clustering.
a) Image to be clustered
b) Manhattan Distance,
c) Euclidean Distance,
d) Standardized Euclidean Distance,
e) Minkowski Distance (p=20),
f) Mahalanobis Distance

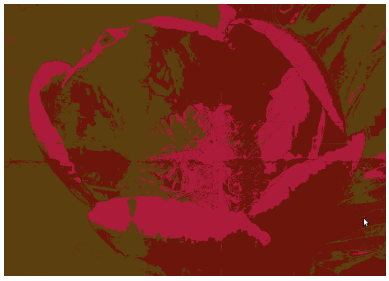

Figure 2: The flower image clustered using a fractional (0.2) p-distance

2.4 Data Separation

Although the curse of dimensionality poses serious problems, processing data with high dimensions also has the advantage that the data are easier to separate. In the majority of cases, N+1 data with N dimensions are linearly separable. In addition, Mercer's theorem (Mercer 1909) can be used, with a mathematical trick (the kernel trick), to describe the data in a potentially infinite dimensional space. This trick is used especially in classification or regression, particularly in SVMs (Vapnik, Golowich, and Smola 1996). However

	Cluster 1	Cluster 2	Cluster 3
Iris-virginica	31	0	19
Iris-setosa	0	50	0
Iris-versicolor	9	0	41
Iris-virginica	48	0	2
Iris-setosa	0	50	0
Iris-versicolor	4	0	46

Table 2: Matching matrices for the iris data clustering. Top: with the Mahalanobis distance; Below: with the standardized Euclidean distance after a kernel PCA

it can also be used to conduct a kernel principal component analysis (Schölkopf, Smola, and Müller 1998). This technique makes it possible to express the data in a higher-dimensional space, in an orthogonal base. Data which are not linearly separable in the initial space become so after being retranscribed in the space created by this technique. The main drawback is that the diagonalization of a $M \times M$ matrix needs to be calculated, where M denotes the number of pieces of data to be clustered. Table 2 presents clustering conducted by a k-means algorithm on the well-known database of UCI iris data using the standardized Euclidean distance of the data as expressed by a linear PCA and a kernel PCA. A k-means algorithm was used and run 3 times, and the best result is presented in this table.

2.5 Summary

We have discussed certain properties of common distances. Table 3 presents the properties of these distances. Mahalanobis distance seems to be an appropriate choice when the dimension number remains reasonable.

Distance	Property
Manhattan	Relatively good data contrast in high dimensions
Euclidean	The barycenter minimizes the sum of the squares of the distances
Standardized Euclidean	The barycenter minimizes the sum of squares of the distances, Uses part of the data distribution
Minkowski $p < 1$	No triangular inequality, Good data contrast in high dimensions
Mahalanobis	The barycenter minimizes the sum of squares of the distances, Uses data correlation

Table 3: Summary of the properties of the most common distances

3 Cluster Similarity

Once the notion of similarity between the data is defined, similarity of data in one cluster (intra-cluster similarity) and similarity between clusters (inter-cluster similarity) must also be clarified. Tables 4 and 5 present the most commonly used inter/intra-cluster distances. Indeed, these metrics are used by algorithms such as hierarchical clustering. Ascending (or agglomerative) hierarchical clustering iteratively groups together clusters with the greatest similarity (inter-cluster similarity). The result of the clustering is strongly influenced by the choice of this metric. But these metrics also serve to evaluate clustering quality. This evaluation can be used as a stopping criteria, or to choose the parameters of the chosen algorithm (such as the number of clusters for a k-means algorithm for example). In this section, we discuss the impact the choice of metric can have on clustering.

Single linkage	$min(d(x,y)), x \in \mathcal{A}, y \in \mathcal{B}$
Complete linkage	$max(d(x,y)), x \in \mathcal{A}, y \in \mathcal{B}$
UPGMA or Average distance	$\frac{1}{\lvert\mathcal{A}\rvert \cdot \lvert\mathcal{B}\rvert} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x,y)$
Average linkage (variation)	$d(\mu_{\mathcal{A}}, \mu_{\mathcal{B}})$ where $\mu_{\mathcal{A}}$ and $\mu_{\mathcal{B}}$ are the arithmetic means of the clusters

Table 4: Common inter-cluster distances

Radius	$max(d(x, \mu_{\mathcal{A}})$ where $\mu_{\mathcal{A}}$ is the arithmetic mean of $\mathcal{A}$
Radius (variation)	$\frac{1}{\lvert\mathcal{A}\rvert} \sum_{x \in \mathcal{A}} d(x, \mu_{\mathcal{A}})$ where $\mu_{\mathcal{A}}$ is the arithmetic mean of $\mathcal{A}$
Diameter	$max(d(x,y)), x \in \mathcal{A}, y \in \mathcal{A}, x \neq y$
Diameter (variation)	$\frac{1}{\lvert\mathcal{A}\rvert \cdot (\lvert\mathcal{A}\rvert - 1)} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{A}} d(x,y)$

Table 5: Common intra-cluster distances

3.1 Outlying Data

In most real problems, the dataset includes outliers. They can be caused by a defective sensor or a typing error for example. The presence of such data, even if there are very few, can greatly influence the inter- and intra-cluster distances. Figure 3 illustrates these variations, with clustering conducted using SLINK (Sibson 1973), CLINK (Defays 1977) and UPGMA. The first row shows that defining the distance by the arithmetic mean of distances between the data of the two groups is much more robust than using the minimum distances of the closest data (SLINK) or the furthest data (CLINK). Indeed, outliers have a tendency to increase intra-cluster distances and decrease inter-cluster distances.

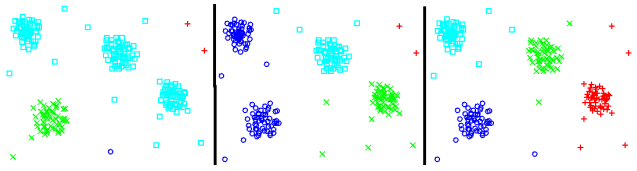

Figure 3: Comparison of common inter-cluster distance for hierarchical clustering on a dataset containing outlying data. The first column corresponds to single linkage, the second to complete linkage and the third to average linkage

3.2 Cluster Shapes

Intra- and inter-cluster distances also influence the shapes of clusters. Of course, the definition of a radius or a diameter for a cluster implies that the cluster is spherical. This hypothesis can be satisfactory for certain problems but not for every case. Figure 4 illustrates the influence of inter-cluster distance of hierarchical data clustering with clusters of various shapes. In this example, only SLINK calculates the clusters satisfactorily.

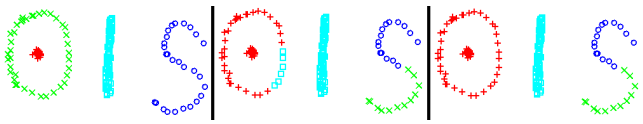

Figure 4: Comparison of common inter-cluster distance for hierarchical clustering with clusters of various shapes. The first column corresponds to single linkage, the second to complete linkage and the third to average linkage

3.3 Size and Density of Heterogeneous Clusters

Figure 5 illustrates the influence of inter-cluster distance of hierarchical clustering of data with clusters of various sizes. It can be seen that CLINK has difficulty clustering this type of data.

The CHAMELEON (Karypis, Han, and Kumar 1999) algorithm offers a measurement of cluster similarity which better accounts for the individuality of the data. It consists to construct a k-nearest neighbors graph of the data and uses the notions of relative inter connectivity and relative closeness of two clusters. These two aspects are defined as functions of the clusters's internal connections and connections between two clusters in the KNN graph. This makes it possible to account for data size and density, for example in order not to systematically group together the small low-density clusters into large, dense clusters.

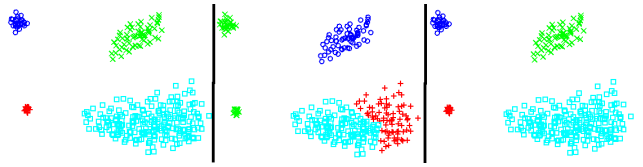

Figure 5: Comparison of common inter-cluster distance for hierarchical clustering with clusters of various sizes. The first column corresponds to single linkage, the second to complete linkage and the third to average linkage

4 Discussion

In this article, we have discussed the different aspects to be taken into account when choosing metrics for data clustering. We have shown how the properties of the most common distances lead to a better support of many specific problems such as size, density, shape of the clusters, or the curse of dimensionality.

These distances are used in most of the clustering algorithms, whether centroids based algorithms as k-means, neural networks such as self-organizing map (Kohonen 1982), Growing Cells Structure or Growing neural gas network (Fritzke 1995), density based approach as DBSCAN (Ester et al. 1996) or OPTICS (Ankerst et al. 1999), or agglomerative hierarchical algorithms. Again, the choice of the best suited algorithm to the problem is not trivial and a thorough study of the characteristics of these different approaches would be helpful. There is indeed no universal clustering algorithm achieving good quality partitioning whatever the problem. In addition, these algorithms generally require parameters to be set correctly. Setting these parameters is a particularly difficult problem of data partitioning, because unlike classification problems, we do not have labels indicating the "solution" which would allow one to perform a cross-validation to adjust settings.

Finally, the difficulty to find the main specificities of a given problem (size, density, linear separation of clusters etc.) raises the issue already extensively discussed on visualization of high-dimensional data.

References

Aggarwal, C. C.; Hinneburg, A.; and Keim, D. A. 2001. On the surprising behavior of distance metrics in high dimensional space. *Database Theory—ICDT 2001* 420–434.

Ankerst, M.; Breunig, M. M.; Kriegel, H.-P.; and Sander, J. 1999. OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD Record*.

Chen, M.-S.; Han, J.; and Yu, P. S. 1996. Data mining: an overview from a database perspective. *Knowledge and Data Engineering, IEEE Transactions on* 8(6):866–883.

Defays, D. 1977. An efficient algorithm for a complete link method. *The Computer Journal* 20(4):364–366.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining*, 226–231.

Fritzke, B. 1995. A growing neural gas network learns topologies. *Advances in neural information processing systems* 7:625–632.

Karypis, G.; Han, E.-H.; and Kumar, V. 1999. Chameleon: hierarchical clustering using dynamic modeling. *Computer* 32(8):68–75.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological cybernetics* 43(1):59–69.

Mercer, J. 1909. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London* 69–70.

Schölkopf, B.; Smola, A.; and Müller, K.-R. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* 10(5):1299–1319.

Sibson, R. 1973. SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal* 16(1):30–34.

Vapnik, V.; Golowich, S. E.; and Smola, A. 1996. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems 9*, 281–287.